Atty. Docket: 03495-0111-01000

UNITED STATES PATENT APPLICATION

OF

BERNARD DUJON, ANDRE CHOULIKA,
LAURENCE COLLEAUX, CECILE FAIRHEAD,
ARRAUD PERRIN, ARNE PLESSIS, AND
AGNES THIERRY

FOR

NUCLEOTIDE SEQUENCE ENCODING THE ENZYME I-*Sce*I AND THE USES THEREOF

CROSS-REFERENCE TO RELATED APPLICATION

This is a continuation-in-part of application Serial No. 07/879,689, filed May 5, 1992, presently pending. The entire disclosure of the prior application is relied upon and incorporated herein by reference.

BACKGROUND OF THE INVENTION

This invention relates to a nucleotide sequence that encodes the restriction endonuclease I-*Sce*I. This invention also relates to vectors containing the nucleotide sequence, cells transformed with the vectors, transgenic animals based on the vectors, and cell lines derived from cells in the animals. This invention also relates to the use of I-*Sce*I for mapping eukaryotic genomes and for in vivo site directed genetic recombination.

The ability to introduce genes into the germ line of mammals is of great interest in biology. The propensity of mammalian cells to take up exogenously added DNA and to express genes included in the DNA has been known for many years. The results of gene manipulation are inherited by the offspring of these animals. All cells of these offspring inherit the introduced gene as part of their genetic make-up. Such animals are said to be transgenic.

Transgenic mammals have provided a means for studying gene regulation during embryogenesis and in differentiation, for studying the action of genes, and for studying the intricate interaction of cells in the immune system. The whole

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
1 202 408 4000

animal is the ultimate assay system for manipulated genes, which direct complex biological processes.

Transgenic animals can provide a general assay for functionally dissecting DNA sequences responsible for tissue specific or developmental regulation of a variety of genes. In addition, transgenic animals provide useful vehicles for expressing recombinant proteins and for generating precise animal models of human genetic disorders.

For a general discussion of gene cloning and expression in animals and animal cells, see Old and Primrose, "Principles of Gene Manipulation," Blackwell Scientific Publications, London (1989), page 255 *et seq.*

Transgenic lines, which have a predisposition to specific diseases and genetic disorders, are of great value in the investigation of the events leading to these states. It is well known that the efficacy of treatment of a genetic disorder may be dependent on identification of the gene defect that is the primary cause of the disorder. The discovery of effective treatments can be expedited by providing an animal model that will lead to the disease or disorder, which will enable the study of the efficacy, safety, and mode of action of treatment protocols, such as genetic recombination.

One of the key issues in understanding genetic recombination is the nature of the initiation step. Studies of homologous recombination in bacteria and fungi have led to the proposal of two types of initiation mechanisms. In the first model, a single-strand nick initiates strand assimilation and branch migration (Meselson and Radding 1975). Alternatively,

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
[illegible] STREET, N.W.
WASHINGTON, D.C. 20005
202 408 4000

a double-strand break may occur, followed by a repair mechanism that uses an uncleaved homologous sequence as a template (Resnick and Martin 1976). This latter model has gained support from the fact that integrative transformation in yeast is dramatically increased when the transforming plasmid is linearized in the region of chromosomal homology (Orr-Weaver, Szostak and Rothstein 1981) and from the direct observation of a double-strand break during mating type interconversion of yeast (Strathern et al. 1982). Recently, double-strand breaks have also been characterized during normal yeast meiotic recombination (Sun et al. 1989; Alani, Padmore and Kleckner 1990).

Several double-strand endonuclease activities have been characterized in yeast: HO and intron encoded endonucleases are associated with homologous recombination functions, while others still have unknown genetic functions (Endo-*Sce*I, Endo-*Sce*II) (Shibata et al. 1984; Morishima et al. 1990). The HO site-specific endonuclease initiates mating-type interconversion by making a double-strand break near the YZ junction of *MAT* (Kostriken et al. 1983). The break is subsequently repaired using the intact *HML* or *HMR* sequences and resulting in ectopic gene conversion. The HO recognition site is a degenerate 24 bp non-symmetrical sequence (Nickoloff, Chen, and Heffron 1986; Nickoloff, Singer and Heffron 1990). This sequence has been used as a "recombinator" in artificial constructs to promote intra- and intermolecular mitotic and meiotic recombination (Nickoloff,

Chen and Heffron, 1986; Kolodkin, Klar and Stahl 1986; Ray et al. 1988, Rudin and Haber, 1988; Rudin, Sugarman, and Haber 1989).

The two-site specific endonucleases, I-*Sce*I (Jacquier and Dujon 1985) and I-*Sce*II (Delahodde et al. 1989; Wenzlau et al. 1989), that are responsible for intron mobility in mitochondria, initiate a gene conversion that resembles the HO-induced conversion (see Dujon 1989 for review). I-*Sce*I, which is encoded by the optional intron Sc LSU.1 of the 21S rRNA gene, initiates a double-strand break at the intron insertion site (Macreadie et al. 1985; Dujon et al. 1985; Colleaux et al. 1986). The recognition site of I-*Sce*I extends over an 18 bp non-symmetrical sequence (Colleaux et al. 1988). Although the two proteins are not obviously related by their structure (HO is 586 amino acids long while I-*Sce*I is 235 amino acids long), they both generate 4 bp staggered cuts with 3'OH overhangs within their respective recognition sites. It has been found that a mitochondrial intron-encoded endonuclease, transcribed in the nucleus and translated in the cytoplasm, generates a double-strand break at a nuclear site. The repair events induced by I-*Sce*I are identical to those initiated by HO.

In summary, there exists a need in the art for reagents and methods for providing transgenic animal models of human diseases and genetic disorders. The reagents can be based on the restriction enzyme I-*Sce*I and the gene encoding this enzyme. In particular, there exists a need for reagents and

methods for replacing a natural gene with another gene that is capable of alleviating the disease or genetic disorder.

SUMMARY OF THE INVENTION

Accordingly, this invention aids in fulfilling these needs in the art. Specifically, this invention relates to an isolated DNA encoding the enzyme I-*Sce*I. The DNA has the following nucleotide sequence:

```
                                                ATG CAT ATG AAA AAC ATC AAA AAA AAC CAG GTA ATG 2670
                                                M   H   M   K   N   I   K   K   N   Q   V   M   12

2671 AAC CTC GGT CCG AAC TCT AAA CTG CTG AAA GAA TAC AAA TCC CAG CTG ATC GAA CTG AAC 2730
  13 N   L   G   P   N   S   K   L   L   K   E   Y   K   S   Q   L   I   E   L   N   32

2731 ATC GAA CAG TTC GAA GCA GGT ATC GGT CTG ATC CTG GGT GAT GCT TAC ATC CGT TCT CGT 2790
  33 I   E   Q   F   E   A   G   I   G   L   I   L   G   D   A   Y   I   R   S   R   52

2791 GAT GAA GGT AAA ACC TAC TGT ATG CAG TTC GAG TGG AAA AAC AAA GCA TAC ATG GAC CAC 2850
  53 D   E   G   K   T   Y   C   M   Q   F   E   W   K   N   K   A   Y   M   D   H   72

2851 GTA TGT CTG CTG TAC GAT CAG TGG GTA CTG TCC CCG CCG CAC AAA AAA GAA CGT GTT AAC 2910
  73 V   C   L   L   Y   D   Q   W   V   L   S   P   P   H   K   K   E   R   V   N   92

2911 CAC CTG GGT AAC CTG GTA ATC ACC TGG GGC GCC CAG ACT TTC AAA CAC CAA GCT TTC AAC 2970
  93 H   L   G   N   L   V   I   T   W   G   A   Q   T   F   K   H   Q   A   F   N  112

2971 AAA CTG GCT AAC CTG TTC ATC GTT AAC AAC AAA AAA ACC ATC CCG AAC AAC CTG GTT GAA 3030
 113 K   L   A   N   L   F   I   V   N   N   K   K   T   I   P   N   N   L   V   E  132

3031 AAC TAC CTG ACC CCG ATG TCT CTG GCA TAC TGG TTC ATG GAT GAT GGT GGT AAA TGG GAT 3090
 133 N   Y   L   T   P   M   S   L   A   Y   W   F   M   D   D   G   G   K   W   D  152

3091 TAC AAC AAA AAC TCT ACC AAC AAA TCG ATC GTA CTG AAC ACC CAG TCT TTC ACT TTC GAA 3150
 153 Y   N   K   N   S   T   N   K   S   I   V   L   N   T   Q   S   F   T   F   E  172

3151 GAA GTA GAA TAC CTG GTT AAG GGT CTG CGT AAC AAA TTC CAA CTG AAC TGT TAC GTA AAA 3210
 173 E   V   E   Y   L   V   K   G   L   R   N   K   F   Q   L   N   C   Y   V   K  192

3211 ATC AAC AAA AAC AAA CCG ATC ATC TAC ATC GAT TCT ATG TCT TAC CTG ATC TTC TAC AAC 3270
 193 I   N   K   N   K   P   I   I   Y   I   D   S   M   S   Y   L   I   F   Y   N  212

3271 CTG ATC AAA CCG TAC CTG ATC CCG CAG ATG ATG TAC AAA CTG CCG AAC ACT ATC TCC TCC 3330
 213 L   I   K   P   Y   L   I   P   Q   M   M   Y   K   L   P   N   T   I   S   S  232

3331 GAA ACT TTC CTG AAA TAA
 233 E   T   F   L   K   *
```

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000

This invention also relates to a DNA sequence comprising a promoter operatively linked to the DNA sequence of the invention encoding the enzyme I-*Sce*I.

This invention further relates to an isolated RNA complementary to the DNA sequence of the invention encoding the enzyme I-*Sce*I and to the other DNA sequences described herein.

In another embodiment of the invention, a vector is provided. The vector comprises a plasmid, bacteriophage, or cosmid vector containing the DNA sequence of the invention encoding the enzyme I-*Sce*I.

In addition, this invention relates to *E. coli* or eukaryotic cells transformed with a vector of the invention.

Also, this invention relates to transgenic animals containing the DNA sequence encoding the enzyme I-*Sce*I and cell lines cultured from cells of the transgenic animals.

In addition, this invention relates to a transgenic organism in which at least one restriction site for the enzyme I-*Sce*I has been inserted in a chromosome of the organism.

Further, this invention relates to a method of genetically mapping a eukaryotic genome using the enzyme I-*Sce*I.

This invention also relates to a method for in vivo site directed recombination in an organism using the enzyme I-*Sce*I.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
1 202 408 4000

7

BRIEF DESCRIPTION OF THE DRAWINGS

This invention will be more fully described with reference to the drawings in which:

Fig. 1 depicts the universal code equivalent of the mitochondrial I-*Sce*I gene.

Fig. 2 depicts the nucleotide sequence of the invention encoding the enzyme I-*Sce*I and the amino acid sequence of the natural I-*Sce*I enzyme.

Fig. 3 depicts the I-*Sce*I recognition sequence and indicates possible base mutations in the recognition site and the effect of such mutations on stringency of recognition.

Fig. 4 is the nucleotide sequence and deduced amino acid sequence of a region of plasmid pSCM525. The nucleotide sequence of the invention encoding the enzyme I-*Sce*I is enclosed in the box.

Fig. 5 depicts variations around the amino acid sequence of the enzyme I-*Sce*I.

Fig. 6 shows Group I intron encoding endonucleases and related endonucleases.

Fig. 7 depicts yeast expression vectors containing the synthetic gene for I-*Sce*I.

Fig. 8 depicts the mammalian expression vector PRSV I-*Sce*I.

Fig. 9 is a restriction map of the plasmid pAF100. (See also YEAST, 6:521-534, 1990, which is relied upon and incorporated by reference herein).

Figs. 10A and 10B show the nucleotide sequence and restriction sites of regions of the plasmid pAF100.

Fig. 11 depicts an insertion vector pTSMω, pTKMω, and pTTcω containing the I-*Sce*I site for *E. coli* and other bacteria.

Fig. 12 depicts an insertion vector pTYW6 containing the I-*Sce*I site for yeast.

Fig. 13 depicts an insertion vector PMLV LTR SAPLZ containing the I-*Sce*I site for mammalian cells.

Fig. 14 depicts a set of seven transgenic yeast strains cleaved by I-*Sce*I. Chromosomes from FY1679 (control) and from seven transgenic yeast strains with I-*Sce*I sites inserted at various positions along chromosome XI were treated with I-*Sce*I. DNA was electrophoresed on 1% agarose (SeaKem) gel in 0.25 X TBE buffer at 130 V and 12°C on a Rotaphor apparatus (Biometra) for 70 hrs using 100 sec to 40 sec decreasing pulse times. (A) DNA was stained with ethidium bromide (0.2μg/ml) and transferred to a Hybond N (Amersham) membrane for hybridization. (B) ^{32}P labelled cosmid pUKG040 which hybridizes with the shortest fragment of the set was used as a probe. Positions of chromosome XI and shorter chromosomes are indicated.

Fig. 15 depicts the rationale of the nested chromosomal fragmentation strategy for genetic mapping. (A) Positions of *I-Sce*I sites are placed on the map, irrespective of the left/right orientation (shorter fragments are arbitrarily placed on the left). Fragment sizes as measured from PFGE (Fig. 14A) are indicated in kb (note that the sum of the two fragment sizes varies slightly due to the limit of precision of each measurement). (B) Hybridization with the probe that

hybridizes the shortest fragment of the set determines the orientation of each fragment (see Fig. 14B). Fragments that hybridize with the probe (full lines) have been placed arbitrarily to the left. (C) Transgenic yeast strains have been ordered with increasing sizes of hybridizing chromosome fragments. (D) Deduced *I-SceI* map with minimal and maximal size of intervals indicated in kb (variations in some intervals are due to limitations of PFGE measurements). (E) Chromosome subfragments are used as probes to assign each cosmid clone to a given map interval or across a given *I-SceI* site.

Fig. 16 depicts mapping of the *I-SceI* sites of transgenic yeast strains by hybridization with left end and right end probes of chromosome XI. Chromosomes from FY1679 (control) and the seven transgenic yeast strains were digested with *I-SceI*. Transgenic strains were placed in order as explained in Fig. 15. Electrophoresis conditions were as in Fig. 14. ^{32}P labelled cosmids pUKG040 and pUKG066 were used as left end and right end probes, respectively.

Fig. 17 depicts mapping of a cosmid collection using the nested chromosomal fragments as probes. Cosmid DNAs were digested with EcoRI and electrophoresed on 0.9% agarose (SeaKem) gel at 1.5 V/cm for 14 hrs, stained with ethidium bromide and transferred to a Hybond N membrane. Cosmids were placed in order from previous hybridizations to help visualize the strategy. Hybridizations were carried out serially on three identical membranes using left end nested chromosome fragments purified on PFGE (see Fig. 16) as

probes. A: ethidium bromide staining (ladder is the BRL "1kb ladder"), B: membrane #1, probe: Left tel to A302 site, C: membrane #1, probe: Left tel to M57 site, D: membrane #2, probe: Left tel to H81 site, E: membrane #2, probe: Left tel to T62 site, F: membrane #3, probe: Left tel to G41 site, G: membrane #3, probe: Left tel to D304 site, H: membrane #3, probe: entire chromosome XI.

Fig. 18 depicts a map of the yeast chromosome XI as determined from the nested chromosomal fragmentation strategy. The chromosome is divided into eight intervals (with sizes indicated in kb, see Fig. 15D) separated by seven *I-SceI* sites (E40, A302 ...). Cosmid clones falling either within intervals or across a given *I-SceI* site are listed below intervals or below interval boundaries, respectively. Cosmid clones that hybridize with selected genes used as probes are indicated by letters (a-i). They localize the gene with respect to the *I-SceI* map and allow comparison with the genetic map (top).

Fig. 19 depicts diagrams of successful site directed homologous recombination experiments performed in yeast.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The genuine mitochondrial gene (ref. 8) cannot be expressed in *E. coli*, yeast or other organisms due to the peculiarities of the mitochondrial genetic code. A "universal code equivalent" has been constructed by *in vitro* site-directed mutagenesis. Its sequence is given in Fig. 1. Note

that all non-universal codons (except two CTN) have been replaced together with some codons extremely rare in *E. coli*.

The universal code equivalent has been successfully expressed in *E. coli* and determines the synthesis of an active enzyme. However, expression levels remained low due to the large number of codons that are extremely rare in *E. coli*. Expression of the "universal code equivalent" has been detected in yeast.

To optimize gene expression in heterologous systems, a synthetic gene has been designed to encode a protein with the genuine amino acid sequence of I-*Sce*I using, for each codon, that most frequently used in *E. coli*. The sequence of the synthetic gene is given in Fig. 2. The synthetic gene was constructed *in vitro* from eight synthetic oligonucleotides with partial overlaps. Oligonucleotides were designed to allow mutual priming for second strand synthesis by Klenow polymerase when annealed by pairs. The elongated pairs were then ligated into plasmids. Appropriately placed restriction sites within the designed sequence allowed final assembly of the synthetic gene by *in vitro* ligation. The synthetic gene has been successfully expressed in both *E. coli* and yeast.

1. <u>I-*Sce*I Gene Sequence</u>

This invention relates to an isolated DNA sequence encoding the enzyme I-*Sce*I. The enzyme I-*Sce*I is an endonuclease. The properties of the enzyme (ref. 14) are as follows:

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000

I-*Sce*I is a double-stranded endonuclease that cleaves DNA within its recognition site. I-*Sce*I generates a 4bp staggered cut with 3'OH overhangs.

Substrate: Acts only on double-stranded DNA. Substrate DNA can be relaxed or negatively supercoiled.

Cations: Enzymatic activity requires Mg^{++} (8 mM is optimum). Mn^{++} can replace Mg^{++}, but this reduces the stringency of recognition.

Optimum conditions for activity: high pH (9 to 10), temperature 20-40°C, no monovalent cations.

Enzyme stability: I-*Sce*I is unstable at room temperature. The enzyme-substrate complex is more stable than the enzyme alone (presence of recognition sites stabilizes the enzyme.)

The enzyme I-*Sce*I has a known recognition site. (ref. 14.) The recognition site of I-*Sce*I is a non-symmetrical sequence that extends over 18 bp as determined by systematic mutational analysis. The sequence reads: (arrows indicate cuts)

5' TAGGGATAACAGGGTAAT 3'

3' ATCCCTATTGTCCCATTA 5'

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, D.C. 20005
(202) 408-4000

The recognition site corresponds, in part, to the upstream exon and, in part, to the downstream exon of the intron plus form of the gene.

The recognition site is partially degenerate: single base substitutions within the 18 bp long sequence result in either complete insensitivity or reduced sensitivity to the enzyme, depending upon position and nature of the substitution.

The stringency of recognition has been measured on:

-1- mutants of the site.

-2- the total yeast genome (*Saccharomyces cerevisiae*, genome complexity is 1.4×10^7 bp). Data are unpublished.

Results are:

-1- Mutants of the site: As shown in Fig. 3, there is a general shifting of stringency, i.e., mutants severely affected in Mg^{++} become partially affected in Mn^{++}, mutants partially affected in Mg^{++} become unaffected in Mn^{++}.

-2- Yeast: In magnesium conditions, no cleavage is observed in normal yeast. In the same condition, DNA from transgenic yeasts is cleaved to completion at the artificially inserted I-*Sce*I site and no other cleavage site can be detected. If magnesium is replaced by manganese, five additional cleavage sites are revealed in the entire yeast genome, none of which is cleaved to completion. Therefore, in manganese the enzyme reveals

an average of 1 site for ca. 3 millions based pairs (5/ 1.4×10^7 bp).

Definition of the recognition site: important bases are indicated in Fig. 3. They correspond to bases for which severely affected mutants exist. Notice however that:

-1- All possible mutations at each position have not been determined; therefore a base that does not correspond to a severely affected mutant may still be important if another mutant was examined at this very same position.

-2- There is no clear-cut limit between a very important base (all mutants are severely affected) and a moderately important base (some of the mutants are severely affected). There is a continuum between excellent substrates and poor substrates for the enzyme.

The expected frequency of natural I-*Sce*I sites in a random DNA sequence is, therefore, equal to $(0.25)^{-18}$ or (1.5×10^{-11}). In other words, one should expect one natural site for the equivalent of ca. 20 human genomes, but the frequency of degenerate sites is more difficult to predict.

I-*Sce*I belongs to a "degenerate" subfamily of the two-dodecapeptide family. Conserved amino acids of the dodecapeptide motifs are required for activity. In particular, the aspartic residues at positions 9 of the two dodecapeptides cannot be replaced, even with

glutamic residues. It is likely that the dodecapeptides form the catalytic site or part of it.

Consistent with the recognition site being non-symmetrical, it is likely that the endonucleolytic activity of I-*Sce*I requires two successive recognition steps: binding of the enzyme to the downstream half of the site (corresponding to the downstream exon) followed by binding of the enzyme to the upstream half of the site (corresponding to the upstream exon). The first binding is strong, the second is weaker, but the two are necessary for cleavage of DNA. *In vitro*, the enzyme can bind the downstream exon alone as well as the intron-exon junction sequence, but no cleavage results.

The evolutionarily conserved dodecapeptide motifs of intron-encoded I-*Sce*I are essential for endonuclease activity. It has been proposed that the role of these motifs is to properly position the acidic amino acids with respect to the DNA sequence recognition domains of the enzyme for the catalysis of phosphodiester bond hydrolysis (ref. P3).

The nucleotide sequence of the invention, which encodes the natural I-*Sce*I enzyme is shown in Fig. 2. The nucleotide sequence of the gene of the invention was derived by dideoxynucleotide sequencing. The base sequences of the nucleotides are written in the 5'——>3' direction. Each of

the letters shown is a conventional designation for the following nucleotides:

A	Adenine
G	Guanine
T	Thymine
C	Cytosine.

It is preferred that the DNA sequence encoding the enzyme I-*Sce*I be in a purified form. For instance, the sequence can be free of human blood-derived proteins, human serum proteins, viral proteins, nucleotide sequences encoding these proteins, human tissue, human tissue components, or combinations of these substances. In addition, it is preferred that the DNA sequence of the invention is free of extraneous proteins and lipids, and adventitious microorganisms, such as bacteria and viruses. The essentially purified and isolated DNA sequence encoding I-*Sce*I is especially useful for preparing expression vectors.

Plasmid pSCM525 is a pUC12 derivative, containing an artificial sequence encoding the DNA sequence of the invention. The nucleotide sequence and deduced amino acid sequence of a region of plasmid pSCM525 is shown in Fig. 4. The nucleotide sequence of the invention encoding I-SceI is enclosed in the box. The artificial gene is a *Bam*HI - *Sal*I piece of DNA sequence of 723 base pairs, chemically synthesized and assembled. It is placed under *tac* promoter control. The DNA sequence of the artificial gene differs from the natural coding sequence or its universal code equivalent described in Cell (1986), Vol. 44, pages 521-533.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, D.C. 20005
(202) 408-4000

However, the translation product of the artificial gene is identical in sequence to the genuine omega-endonuclease except for the addition of a Met-His at the N-terminus. It will be understood that this modified endonuclease is within the scope of this invention.

Plasmid pSCM525 can be used to transform any suitable *E. coli* strain and transformed cells become ampicillin-resistant. Synthesis of the omega-endonuclease is obtained by addition of I.P.T.G. or an equivalent inducer of the lactose operon system.

A plasmid identified as pSCM525 containing the enzyme I-*Sce*I was deposited in *E. coli* strain TG1 with the Collection Nationale de Cultures de Microorganismes (C.N.C.M.) of Institut Pasteur in Paris, France on November 22, 1990, under culture collection deposit Accession No. I-1014. The nucleotide sequence of the invention is thus available from this deposit.

The gene of the invention can also be prepared by the formation of 3′——>5′ phosphate linkages between nucleoside units using conventional chemical synthesis techniques. For example, the well-known phosphodiester, phosphotriester, and phosphite triester techniques, as well as known modifications of these approaches, can be employed. Deoxyribonucleotides can be prepared with automatic synthesis machines, such as those based on the phosphoramidite approach. Oligo- and polyribonucleotides can also be obtained with the aid of RNA ligase using conventional techniques.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000

This invention of course includes variants of the DNA sequence of the invention exhibiting substantially the same properties as the sequence of the invention. By this it is meant that DNA sequences need not be identical to the sequence disclosed herein. Variations can be attributable to single or multiple base substitutions, deletions, or insertions or local mutations involving one or more nucleotides not substantially detracting from the properties of the DNA sequence as encoding an enzyme having the cleavage properties of the enzyme I-*Sce*I.

Fig. 5 depicts some of the variations that can be made around the I-*Sce*I amino acid sequence. It has been demonstrated that the following positions can be changed without affecting enzyme activity:

positions -1 and -2 are not natural. The two amino acids are added due to cloning strategies.
positions 1 to 10: can be deleted.
position 36: G is tolerated.
position 40: M or V are tolerated.
position 41: S or N are tolerated.
position 43: A is tolerated.
position 46: V or N are tolerated.
position 91: A is tolerated.
positions 123 and 156: L is tolerated.
position 223: A and S are tolerated.

It will be understood that enzymes containing these modifications are within the scope of this invention.

Changes to the amino acid sequence in Fig. 5 that have been demonstrated to affect enzyme activity are as follows:

position 19: L to S
position 38: I to S or N
position 39: G to D or R
position 40: L to Q
position 42: L to R

position 44: D to E, G or H
position 45: A to E or D
position 46: Y to D
position 47: I to R or N
position 80: L to S
position 144: D to E
position 145: D to E
position 146: G to E
position 147: G to S

It will also be understood that the present invention is intended to encompass fragments of the DNA sequence of the invention in purified form, where the fragments are capable of encoding enzymatically active I-*Sce*I.

The DNA sequence of the invention coding for the enzyme I-*Sce*I can be amplified in the well known polymerase chain reaction (PCR), which is useful for amplifying all or specific regions of the gene. See e.g., S. Kwok et al., J. Virol., 61:1690-1694 (1987); U.S. Patent 4,683,202; and U.S. Patent 4,683,195. More particularly, DNA primer pairs of known sequence positioned 10-300 base pairs apart that are complementary to the plus and minus strands of the DNA to be amplified can be prepared by well known techniques for the synthesis of oligonucleotides. One end of each primer can be extended and modified to create restriction endonuclease sites when the primer is annealed to the DNA. The PCR reaction mixture can contain the DNA, the DNA primer pairs, four deoxyribonucleoside triphosphates, $MgCl_2$, DNA polymerase, and conventional buffers. The DNA can be amplified for a number of cycles. It is generally possible to increase the sensitivity of detection by using a multiplicity of cycles, each cycle consisting of a short period of denaturation of the DNA at an elevated temperature, cooling of the reaction mixture,

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000

and polymerization with the DNA polymerase. Amplified sequences can be detected by the use of a technique termed oligomer restriction (OR). See, R. K. Saiki et al., Bio/Technology 3:1008-1012 (1985).

The enzyme I-*Sce*I is one of a number of endonucleases with similar properties. Following is a listing of related enzymes and their sources.

Group I intron encoded endonucleases and related enzymes are listed below with references. Recognition sites are shown in Fig. 6.

Enzyme	Encoded by	Ref
I-*Sce*I	Sc LSU-1 intron	this work
I-*Sce*II	Sc cox1-4 intron	Sargueil et al., NAR (1990) 18, 5659-5665
I-*Sce*III	Sc cox1-3 intron	Sargueil et al., MGG (1991) 225, 340-341
I-*Sce*IV	Sc cox1-5a intron	Seraphin et al. (1992) in press
I-*Ceu*I	Ce LSU-5 intron	Marshall, Lemieux Gene (1991) 104, 241-245
I-*Cre*I	Cr LSU-1 intron	Rochaix (unpublished)
I-*Ppo*I	Pp LSU-3 intron	Muscarella et al., MCB (1990) 10, 3386-3396
I-*Tev*I	T4 td-1 intron	Chu et al., PNAS (1990) 87, 3574-3578 and Bell-Pedersen et al. NAR (1990) 18, 3763-3770.
I-*Tev*II	T4 sunY intron	Bell-Pedersen et al. NAR (1990) 18, 3763-3770.
I-*Tev*III	RB3 nrdB-1 intron	Eddy, Gold, Genes Dev. (1991) 5, 1032-1041
HO	HO yeast gene	Nickoloff et al., MCB (1990) 10, 1174-1179
Endo *Sce*I	RF3 yeast mito. gene	Kawasaki et al., JBC (1991) 266, 5342-5347

Putative new enzymes (genetic evidence but no activity as yet) are I-*Csm*I from cytochrome b intron 1 of *Chlamydomonas smithii* mitochondria (ref. 15), I-*Pan*I from cytochrome b intron 3 of *Podospora anserina* mitochondria (Jill Salvo), and

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, D.C. 20005
(202) 408-4000

probably enzymes encoded by introns Nc nd1·1 and Nc cob·! from *Neurospora crassa*.

The I-endonucleases can be classified as follows:

Class I: Two dodecapeptide motifs, 4 bp staggered cut with 3' OH overhangs, cut internal to recognition site

Subclass "I-*SceI*"	Other subclasses
I-*SceI*	I-*SceII*
I-*SceIV*	I-*SceIII*
I-*CsmI*	I-*CeuI* (only one dodecapeptide motif)
I-*PanI*	I-*CreI* (only one dodecapeptide motif)
	HO
	TFP1-408 (HO homolog)
	Endo *SceI*

Class II: GIY-(N_{10-11}) YIG motif, 2 bp staggered cut with 3' OH overhangs, cut external to recognition site:

I-*TevI*

Class III: no typical structural motifs, 4 bp staggered cut with 3' OH overhangs, cut internal to recognition site:

I-*PpoI*

Class IV: no typical structural motifs, 2 bp staggered cut with 3' OH overhangs, cut external to recognition site:

I-*TevII*

Class V: no typical structural motifs, 2 bp staggered cut with 5' OH overhangs:

I-*TevIII*.

T0220

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
(202) 408-4000

2. Nucleotide Probes Containing the I-*SceI* Gene of The Invention

The DNA sequence of the invention coding for the enzyme I-*SceI* can also be used as a probe for the detection of a nucleotide sequence in a biological material, such as tissue or body fluids. The probe can be labeled with an atom or inorganic radical, most commonly using a radionuclide, but also perhaps with a heavy metal. Radioactive labels include ^{32}P, ^{3}H, ^{14}C, or the like. Any radioactive label can be employed, which provides for an adequate signal and has sufficient half-life. Other labels include ligands that can serve as a specific binding member to a labeled antibody, fluorescers, chemiluminescers, enzymes, antibodies which can serve as a specific binding pair member for a labeled ligand, and the like. The choice of the label will be governed by the effect of the label on the rate of hybridization and binding of the probe to the DNA or RNA. It will be necessary that the label provide sufficient sensitivity to detect the amount of DNA or RNA available for hybridization.

When the nucleotide sequence of the invention is used as a probe for hybridizing to a gene, the nucleotide sequence is preferably affixed to a water insoluble solid, porous support, such as nitrocellulose paper. Hybridization can be carried out using labeled polynucleotides of the invention and conventional hybridization reagents. The particular hybridization technique is not essential to the invention.

The amount of labeled probe present in the hybridization solution will vary widely, depending upon the nature of the

label, the amount of the labeled probe which can reasonably bind to the support, and the stringency of the hybridization. Generally, substantial excesses of the probe over stoichiometric will be employed to enhance the rate of binding of the probe to the fixed DNA.

Various degrees of stringency of hybridization can be employed. The more severe the conditions, the greater the complementarity that is required for hybridization between the probe and the polynucleotide for duplex formation. Severity can be controlled by temperature, probe concentration, probe length, ionic strength, time, and the like. Conveniently, the stringency of hybridization is varied by changing the polarity of the reactant solution. Temperatures to be employed can be empirically determined or determined from well known formulas developed for this purpose.

3. Nucleotide Sequences Containing the Nucleotide Sequence Encoding I-SceI

This invention also relates to the DNA sequence of the invention encoding the enzyme I-*SceI*, wherein the nucleotide sequence is linked to other nucleic acids. The nucleic acid can be obtained from any source, for example, from plasmids, from cloned DNA or RNA, or from natural DNA or RNA from any source, including prokaryotic and eukaryotic organisms. DNA or RNA can be extracted from a biological material, such as biological fluids or tissue, by a variety of techniques including those described by Maniatis et al., Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, New

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
(202) 408-4000

York (1982). The nucleic acid will generally be obtained from a bacteria, yeast, virus, or a higher organism, such as a plant or animal. The nucleic acid can be a fraction of a more complex mixture, such as a portion of a gene contained in whole human DNA or a portion of a nucleic acid sequence of a particular microorganism. The nucleic acid can be a fraction of a larger molecule or the nucleic acid can constitute an entire gene or assembly of genes. The DNA can be in a single-stranded or double-stranded form. If the fragment is in single-stranded form, it can be converted to double-stranded form using DNA polymerase according to conventional techniques.

The DNA sequence of the invention can be linked to a structural gene. As used herein, the term "structural gene" refers to a DNA sequence that encodes through its template or messenger mRNA a sequence of amino acids characteristic of a specific protein or polypeptide. The nucleotide sequence of the invention can function with an expression control sequence, that is, a DNA sequence that controls and regulates expression of the gene when operatively linked to the gene.

4. Vectors Containing the Nucleotide Sequence of the Invention

This invention also relates to cloning and expression vectors containing the DNA sequence of the invention coding for the enzyme I-*Sce*I.

More particularly, the DNA sequence encoding the enzyme can be ligated to a vehicle for cloning the sequence. The

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, D.C. 20005
(202) 408-4000

major steps involved in gene cloning comprise procedures for separating DNA containing the gene of interest from prokaryotes or eukaryotes, cutting the resulting DNA fragment and the DNA from a cloning vehicle at specific sites, mixing the two DNA fragments together, and ligating the fragments to yield a recombinant DNA molecule. The recombinant molecule can then be transferred into a host cell, and the cells allowed to replicate to produce identical cells containing clones of the original DNA sequence.

The vehicle employed in this invention can be any double-stranded DNA molecule capable of transporting the nucleotide sequence of the invention into a host cell and capable of replicating within the cell. More particularly, the vehicle must contain at least one DNA sequence that can act as the origin of replication in the host cell. In addition, the vehicle must contain two or more sites for insertion of the DNA sequence encoding the gene of the invention. These sites will ordinarily correspond to restriction enzyme sites at which cohesive ends can be formed, and which are complementary to the cohesive ends on the promoter sequence to be ligated to the vehicle. In general, this invention can be carried out with plasmid, bacteriophage, or cosmid vehicles having these characteristics.

The nucleotide sequence of the invention can have cohesive ends compatible with any combination of sites in the vehicle. Alternatively, the sequence can have one or more blunt ends that can be ligated to corresponding blunt ends in the cloning sites of the vehicle. The nucleotide sequence to

be ligated can be further processed, if desired, by successive exonuclease deletion, such as with the enzyme Bal 31. In the event that the nucleotide sequence of the invention does not contain a desired combination of cohesive ends, the sequence can be modified by adding a linker, an adaptor, or homopolymer tailing.

It is preferred that plasmids used for cloning nucleotide sequences of the invention carry one or more genes responsible for a useful characteristic, such as a selectable marker, displayed by the host cell. In a preferred strategy, plasmids having genes for resistance to two different drugs are chosen. For example, insertion of the DNA sequence into a gene for an antibiotic inactivates the gene and destroys drug resistance. The second drug resistance gene is not affected when cells are transformed with the recombinants, and colonies containing the gene of interest can be selected by resistance to the second drug and susceptibility to the first drug. Preferred antibiotic markers are genes imparting chloramphenicol, ampicillin, or tetracycline resistance to the host cell.

A variety of restriction enzymes can be used to cut the vehicle. The identity of the restriction enzyme will generally depend upon the identity of the ends on the DNA sequence to be ligated and the restriction sites in the vehicle. The restriction enzyme is matched to the restriction sites in the vehicle, which in turn is matched to the ends on the nucleic acid fragment being ligated.

The ligation reaction can be set up using well known techniques and conventional reagents. Ligation is carried out with a DNA ligase that catalyzes the formation of phosphodiester bonds between adjacent 5'-phosphate and the free 3'-hydroxy groups in DNA duplexes. The DNA ligase can be derived from a variety of microorganisms. The preferred DNA ligases are enzymes from *E. coli* and bacteriophage T4. T4 DNA ligase can ligate DNA fragments with blunt or sticky ends, such as those generated by restriction enzyme digestion. *E. coli* DNA ligase can be used to catalyze the formation of phosphodiester bonds between the termini of duplex DNA molecules containing cohesive ends.

Cloning can be carried out in prokaryotic or eukaryotic cells. The host for replicating the cloning vehicle will of course be one that is compatible with the vehicle and in which the vehicle can replicate. When a plasmid is employed, the plasmid can be derived from bacteria or some other organism or the plasmid can be synthetically prepared. The plasmid can replicate independently of the host cell chromosome or an integrative plasmid (episome) can be employed. The plasmid can make use of the DNA replicative enzymes of the host cell in order to replicate or the plasmid can carry genes that code for the enzymes required for plasmid replication. A number of different plasmids can be employed in practicing this invention.

The DNA sequence of the invention encoding the enzyme I-*Sce*I can also be ligated to a vehicle to form an expression vector. The vehicle employed in this case is one in which it

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000

is possible to express the gene operatively linked to a promoter in an appropriate host cell. It is preferable to employ a vehicle known for use in expressing genes in *E. coli*, yeast, or mammalian cells. These vehicles include, for example, the following *E. coli* expression vectors:

pSCM525, which is an *E. coli* expression vector derived from pUC12 by insertion of a tac promoter and the synthetic gene for I-*Sce*I. Expression is induced by IPTG.

pGEXω6, which is an *E. coli* expression vector derived from pGEX in which the synthetic gene from pSCM525 for I-*Sce*I is fused with the glutathione S transferase gene, producing a hybrid protein. The hybrid protein possesses the endonuclease activity.

pDIC73, which is an *E. coli* expression vector derived from pET-3C by insertion of the synthetic gene for I-*Sce*I (*Nde*I - *Bam*HI fragment of pSCM525) under T7 promoter control. This vector is used in strain BL21 (DE3) which expresses the T7 RNA polymerase under IPTG induction.

pSCM351, which is an *E. coli* expression vector derived from pUR291 in which the synthetic gene for I-*Sce*I is fused with the Lac Z gene, producing a hybrid protein.

pSCM353, which is an *E. coli* expression vector derived from pEX1 in which the synthetic gene for I-*Sce*I is fused with the Cro/Lac Z gene, producing a hybrid protein.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, D.C. 20005
(202) 408-4000

Examples of yeast expression vectors are:

pPEX7, which is a yeast expression vector derived from pRP51-Bam O (a LEU2d derivative of pLG-SD5) by insertion of the synthetic gene under the control of the galactose promoter. Expression is induced by galactose.

pPEX408, which is a yeast expression vector derived from pLG-SD5 by insertion of the synthetic gene under the control of the galactose promoter. Expression is induced by galactose.

Several yeast expression vectors are depicted in Fig. 7.

Typical mammalian expression vectors are:

pRSV I-*Sce*I, which is a pRSV derivative in which the synthetic gene (*Bam*HI - *Pst*I fragment from pSCM525) is under the control of the LTR promoter of Rous Sarcoma Virus. This expression vector is depicted in Fig. 8.

Vectors for expression in Chinese Hamster Ovary (CHO) cells can also be employed.

5. Cells Transformed with Vectors of the Invention

The vectors of the invention can be inserted into host organisms using conventional techniques. For example, the vectors can be inserted by transformation, transfection, electroporation, microinjection, or by means of liposomes (lipofection).

Cloning can be carried out in prokaryotic or eukaryotic cells. The host for replicating the cloning vehicle will of course be one that is compatible with the vehicle and in

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
300 I STREET, N.W.
WASHINGTON, D.C. 20005
(202) 408-4000

which the vehicle can replicate. Cloning is preferably carried out in bacterial or yeast cells, although cells of fungal, animal, and plant origin can also be employed. The preferred host cells for conducting cloning work are bacterial cells, such as *E. coli*. The use of *E. coli* cells is particularly preferred because most cloning vehicles, such as bacterial plasmids and bacteriophages, replicate in these cells.

In a preferred embodiment of this invention, an expression vector containing the DNA sequence encoding the nucleotide sequence of the invention operatively linked to a promoter is inserted into a mammalian cell using conventional techniques.

Application of I-*Sce*I for large scale mapping

1. Occurrence of natural sites in various genomes

Using the purified I-*Sce*I enzyme, the occurrence of natural or degenerate sites has been examined on the complete genomes of several species. No natural site was found in *Saccharomyces cerevisiae, Bacillus anthracis, Borrelia burgdorferi, Leptospira biflexa* and *L. interrogans*. One degenerate site was found on T7 phage DNA.

2. Insertion of artificial sites

Given the absence of natural I-*Sce*I sites, artificial sites can be introduced by transformation or transfection. Two cases need to be distinguished: site-directed integration by homologous recombination and random integration by

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
1 202 408 4000

non-homologous recombination, transposon movement or retroviral infection. The first is easy in the case of yeast and a few bacterial species, more difficult for higher eucaryotes. The second is possible in all systems.

3. Insertion vectors

Two types can be distinguished:

-1- Site specific cassettes that introduce the *I-SceI* site together with a selectable marker.

For yeast: all are pAF100 derivatives (Thierry et al. (1990) YEAST 6:521-534) containing the following marker genes:

pAF101: URA3 (inserted in the *Hind*III site)

pAF103: Neo^R (inserted in *Bgl*II site)

pAF104: HIS3 (inserted in *Bgl*II site)

pAF105: Kan^R (inserted in *Bgl*II site)

pAF106: Kan^R (inserted in *Bgl*II site)

pAF107: LYS2 (inserted between *Hind*III and *Eco*R V)

A restriction map of the plasmid pAF100 is shown in Fig. 9. The nucleotide sequence and restriction sites of regions of plasmid pAF100 are shown in Figs. 10A and 10B.

Many transgenic yeast strains with the I-*Sce*I site at various and known places along chromosomes are available.

-2- Vectors derived from transposable elements or retroviruses.

For *E. coli* and other bacteria: mini Tn5 derivatives containing the I-*Sce*I site and

pTSm ω Str^R

pTKm ω Kan^R (See Fig. 11)

pTTc ω Tet^R

For yeast: pTyω6 is a pD123 derivative in which the I-*Sce*I site has been inserted in the LTR of the Ty element. (Fig. 12)

For mammalian cells:

pMLV LTR SAPLZ: containing the I-*Sce*I site in the LTR of MLV and Phleo-LacZ (Fig. 13). This vector is first grown in Ψ2 cells (3T3 derivative, from R. Mulligan). Two transgenic cell lines with the I-*Sce*I site at undetermined locations in the genome are available: 1009 (pluripotent nerve cells, J.F. Nicolas) and D3 (ES cells able to generate transgenic animals).

4. <u>The nested chromosomal fragmentation strategy</u>

The nested chromosomal fragmentation strategy for genetically mapping a eukaryotic genome exploits the unique properties of the restriction endonuclease *I-Sce*I, such as an 18 bp long recognition site. The absence of natural I-*Sce*I recognition sites in most eukaryotic genomes is also exploited in this mapping strategy.

First, one or more I-*Sce*I recognition sites are artificially inserted at various positions in a genome, by

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
1 202 408 4000

homologous recombination using specific cassettes containing selectable markers or by random insertion, as discussed supra. The genome of the resulting transgenic strain is then cleaved completely at the artificially inserted I-*Sce*I site(s) upon incubation with the I-SceI restriction enzyme. The cleavage produces nested chromosomal fragments.

The chromosomal fragments are then purified and separated by pulsed field gel (PFG) electrophoresis, allowing one to "map" the position of the inserted site in the chromosome. If total DNA is cleaved with the restriction enzyme, each artificially introduced I-*Sce*I site provides a unique "molecular milestone" in the genome. Thus, a set of transgenic strains, each carrying a single I-*Sce*I site, can be created which defines physical genomic intervals between the milestones. Consequently, an entire genome, a chromosome or any segment of interest can be mapped using artificially introduced I-SceI restriction sites.

The nested chromosomal fragments may be transferred to a solid membrane and hybridized to a labelled probe containing DNA complementary to the DNA of the fragments. Based on the hybridization banding patterns that are observed, the eukaryotic genome may be mapped. The set of transgenic strains with appropriate "milestones" is used as a reference to map any new gene or clone by direct hybridization.

Example 1: Application of the Nested Chromosomal Fragmentation Strategy to the Mapping of Yeast Chromosome XI

This strategy has been applied to the mapping of yeast chromosome XI of *Saccharamyces cerevisiae*. The I-*Sce*I site was inserted at 7 different locations along chromosome XI of the diploid strain FY1679, hence defining eight physical intervals in that chromosome. Sites were inserted from a URA3-1-I-*Sce*I cassette by homologous recombination. Two sites were inserted within genetically defined genes, TIF1 and FAS1, the others were inserted at unknown positions in the chromosome from five non-overlapping cosmids of our library, taken at random. Agarose embedded DNA of each of the seven transgenic strains was then digested with I-*Sce*I and analyzed by pulsed field gel electrophoresis (Fig. 14A). The position of the I-*Sce*I site of each transgenic strain in chromosome XI is first deduced from the fragment sizes without consideration of the left/right orientation of the fragments. Orientation was determined as follows. The most telomere proximal I-*Sce*I site from this set of strains is in the transgenic E40 because the 50 kb fragment is the shortest of all fragments (Fig. 15A). Therefore, the cosmid clone pUKG040, which was used to insert the I-*Sce*I site in the transgenic E40, is now used as a probe against all chromosome fragments (Fig. 14B). As expected, pUKG040 lights up the two fragments from strain E40 (50 kb and 630 kb, respectively). The large fragment is close to the entire chromosome XI and shows a weak hybridization signal due to the fact that the insert of pUKG040, which is 38 kb long, contains less than 4

kb within the large chromosome fragment. Note that the entire chromosome XI remains visible after I-*Sce*I digestion, due to the fact that the transgenic strains are diploids in which the I-*Sce*I site is inserted in only one of the two homologs. Now, the pUKG040 probe hybridizes to only one fragment of all other transgenic strains allowing unambiguous left/right orientation of I-*Sce*I sites (See Fig. 15B). No significant cross hybridization between the cosmid vector and the chromosome subfragment containing the I-*Sce*I site insertion vector is visible. Transgenic strains can now be ordered such that I-*Sce*I sites are located at increasing distances from the hybridizing end of the chromosome (Fig. 15C) and the I-*Sce*I map can be deduced (Fig. 15D). Precision of the mapping depends upon PFGE resolution and optimal calibration. Note that actual left/right orientation of the chromosome with respect to the genetic map is not known at this step. To help visualize our strategy and to obtain more precise measurements of the interval sizes between I-*Sce*I sites between I-*Sce*I, a new pulsed field gel electrophoresis with the same transgenic strains now placed in order was made (Fig. 16). After transfer, the fragments were hybridized successively with cosmids pUKG040 and pUKG066 which light up, respectively, all fragments from the opposite ends of the chromosome (clone pUKG066 defines the right end of the chromosome as defined from the genetic map because it contains the SIR1 gene. A regular stepwise progression of chromosome fragment sizes is observed. Note some cross

LAW OFFICES
INNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, DC 20005
1 202 408 4000

hybridization between the probe pUKG066 and chromosome III, probably due to some repetitive DNA sequences.

All chromosome fragments, taken together, now define physical intervals as indicated in Fig. 15d. The I-*Sce*I map obtained has an 80 kb average resolution.

Example 2: Application of the Nested Chromosomal Fragmentation Strategy to the Mapping of Yeast Artificial Chromosome (YAC) Clones

This strategy can be applied to YAC mapping with two possibilities.

-1- insertion of the I-*Sce*I site within the gene of interest using homologous recombination in yeast. This permits mapping of that gene in the YAC insert by I-*Sce*I digestion *in vitro*. This has been done and works.

-2- random integration of I-*Sce*I sites along the YAC insert by homologous recombination in yeast using highly repetitive sequences (e.g., B2 in mouse or Alu in human). Transgenic strains are then used as described in ref. P1 to sort libraries or map genes.

The procedure has now been extended to YAC containing 450 kb of Mouse DNA. To this end, a repeated sequence of mouse DNA (called B2) has been inserted in a plasmid containing the I-*Sce*I site and a selectable yeast marker (LYS2). Transformation of the yeast cells containing the recombinant YAC with the plasmid linearized within the B2 sequence resulted in the integration of the I-*Sce*I site at five different locations distributed along the mouse DNA insert. Cleavage at the inserted I-*Sce*I sites using the

enzyme has been successful, producing nested fragments that can be purified after electrophoresis. Subsequent steps of the protocol exactly parallels the procedure described in Example 1.

Example 3: Application of Nested Chromosomal Fragments to the Direct Sorting of Cosmid Libraries

The nested, chromosomal fragments can be purified from preparative PFG and used as probes against clones from a chromosome X1 specific sublibrary. This sublibrary is composed of 138 cosmid clones (corresponding to eight times coverage) which have been previously sorted from our complete yeast genomic libraries by colony hybridization with PFG purified chromosome X1. This collection of unordered clones has been sequentially hybridized with chromosome fragments taken in order of increasing sizes from the left end of the chromosome. Localization of each cosmid clone on the I-*Sce*I map could be unambiguously determined from such hybridizations. To further verify the results and to provide a more precise map, a subset of all cosmid clones, now placed in order, have been digested with EcoRI, electrophoresed and hybridized with the nested series of chromosome fragments in order of increasing sizes from the left end of the chromosome. Results are given in Figure 17.

For a given probe, two cases can be distinguished: cosmid clones in which all EcoRI fragments hybridize with the probe and cosmid clones in which only some of the EcoRI fragments hybridize (i.e., compare pEKG100 to pEKG098 in Fig.

17b). The first category corresponds to clones in which the insert is entirely included in one of the two chromosome fragments, the second to clones in which the insert overlaps an I-*Sce*I site. Note that, for clones of the pEKG series, the EcoRI fragment of 8 kb is entirely composed of vector sequences (pWE15) that do not hybridize with the chromosome fragments. In the case where the chromosome fragment possesses the integration vector, a weak cross hybridization with the cosmid is observed (Fig. 17e).

Examination of Fig. 17 shows that the cosmid clones can unambiguously be ordered with respect to the I-*Sce*I map (Fig. 13E), each clone falling either in a defined interval or across an I-*Sce*I site. In addition, clones from the second category allow us to place some EcoRI fragments on the I-*Sce*I maps, while others remain unordered. The complete set of chromosome XI- specific cosmid clones, covering altogether eight times the equivalent of the chromosome, has been sorted with respect to the I-*Sce*I map, as shown in Fig. 18.

5. Partial restriction mapping using I-*Sce*I

In this embodiment, complete digestion of the DNA at the artificially inserted I-*Sce*I site is followed by partial digestion with bacterial restriction endonucleases of choice. The restriction fragments are then separated by electrophoresis and blotted. Indirect end labelling is accomplished using left or right I-*Sce* half sites. This technique has been successful with yeast chromosomes and should be applicable without difficulty for YAC.

Partial restriction mapping has been done on yeast DNA and on mammalian cell DNA using the commercial enzyme I-*Sce*I. DNA from cells containing an artificially inserted I-*Sce*I site is first cleaved to completion by I-*Sce*I. The DNA is then treated under partial cleavage conditions with bacterial restriction endonucleases of interest (e.g., *Bam*HI) and electrophoresed along with size calibration markers. The DNA is transferred to a membrane and hybridized successively using the short sequences flanking the I-*Sce*I sites on either side (these sequences are known because they are part of the original insertion vector that was used to introduce the I-*Sce*I site). Autoradiography (or other equivalent detection system using non radioactive probes) permit the visualization of ladders, which directly represent the succession of the bacterial restriction endonuclease sites from the I-*Sce*I site. The size of each band of the ladder is used to calculate the physical distance between the successive bacterial restriction endonuclease sites.

Application of I-*Sce*I for *In Vivo* Site Directed Recombination

1. Expression of I-*Sce*I in yeast

The synthetic I-*Sce*I gene has been placed under the control of a galactose inducible promoter on multicopy plasmids pPEX7 and pPEX408. Expression is correct and induces effects on site as indicated below. A transgenic yeast with the I-*Sce*I synthetic gene inserted in a chromosome

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N. W.
WASHINGTON, D.C. 20005
(202) 408-4000

under the control of an inducible promoter can be constructed.

2. Effects of site specific double strand breaks in yeast (refs. 18 and P4)

Effects on plasmid-borne I-*Sce*I sites:

Intramolecular effects are described in detail in Ref. 18. Intermolecular (plasmid to chromosome) recombination can be predicted.

Effects on chromosome integrated I-*Sce*I sites

In a haploid cell, a single break within a chromosome at an artificial I-*Sce*I site results in cell division arrest followed by death (only a few % of survival). Presence of an intact sequence homologous to the cut site results in repair and 100% cell survival. In a diploid cell, a single break within a chromosome at an artificial I-*Sce*I site results in repair using the chromosome homolog and 100% cell survival. In both cases, repair of the induced double strand break results in loss of heterozygosity with deletion of the non homologous sequences flanking the cut and insertion of the non homologous sequences from the donor DNA molecule.

3. Application for *in vivo* recombination YACs in Yeast

Construction of a YAC vector with the I-*Sce*I restriction site next to the cloning site should permit one to induce homologous recombination with another YAC if inserts are

partially overlapping. This is useful for the construction of contigs.

4. Prospects for other organisms

Insertion of an I-*Sce*I restriction site has been done for bacteria (*E. coli*, *Yersinia entorocolitica*, *Y. pestis*, *Y. pseudotuberculosis*), and mouse cells. Cleavage at the artificial I-*Sce*I site *in vitro* has been successful with DNA from the transgenic mouse cells. Expression of I-*Sce*I from the synthetic gene in mammalian or plant cells should be successful.

The I-*Sce*I site has been introduced in mouse cells and bacterial cells as follows:

-1- Mouse cells:

-a- Mouse cells (ψ2) were transfected with the DNA of the vector pMLV LTR SAPLZ containing the I-*Sce*I site using standard calcium phosphate transfection technique.

-b- Transfected cells were selected in DMEM medium containing phleomycin with 5% fetal calf serum and grown under 12% CO_2, 100% humidity at 37°C until they form colonies.

-c- Phleomycin resistant colonies were subcloned once in the same medium.

-d- Clone MLOP014, which gave a titer of 10^5 virus particles per ml, was chosen. This clone was deposited at C.N.C.M. on May 5, 1992 under culture collection accession No. I-1207.

-e- The supernatant of this clone was used to infect other mouse cells (1009) by spreading 10^5 virus particles on

10^5 cells in DMEM medium with 10% fetal calf serum and 5 mg/ml of "polybrain". Medium was replaced 6 hours after infection by the same fresh medium.

-f- 24 hours after infection, phleomycin resistant cells were selected in the same medium as above.

-g- phleomycin resistant colonies were subcloned once in the same medium.

-h- one clone was picked and analyzed. DNA was purified with standard procedures and digested with I-*Sce*I under optimal conditions.

-2- Bacterial cells:

Mini Tn 5 transposons containing the I-*Sce*I recognition site were constructed in *E. coli* by standard recombinant DNA procedures. The mini Tn 5 transposons are carried on a conjugative plasmid. Bacterial conjugation between *E. coli* and *Yersinia* is used to integrate the mini Tn 5 transposon in *Yersinia*. *Yersinia* cells resistant to *Kanamycin, Streptomycin* or *tetracycline* are selected (vectors pTKM-ω, pTSM-ω and pTTc-ω, respectively).

Several strategies can be attempted for the site specific insertion of a DNA fragment from a plasmid into a chromosome. This will make it possible to insert transgenes at predetermined sites without laborious screening steps. Strategies are:

-1- Construction of a transgenic cell in which the I-*Sce*I recognition site is inserted at a unique location in a chromosome. Cotransformation of the transgenic cell with the expression vector and a plasmid containing the gene of

interest and a segment homologous to the sequence in which the I-*Sce*I site is inserted.

-2- Insertion of the I-*Sce*I recognition site next to or within the gene of interest carried on a plasmid. Cotransformation of a normal cell with the expression vector carrying the synthetic I-*Sce*I gene and the plasmid containing the I-*Sce*I recognition site.

-3- Construction of a stable transgenic cell line in which the I-*Sce*I gene has been integrated in the genome under the control of an inducible or constitutive cellular promoter. Transformation of the cell line by a plasmid containing the I-*Sce*I site next to or within the gene of interest.

Site directed homologous recombination: diagrams of successful experiments performed in yeast are given in Fig. 19.

Publications Cited In Application

1. B. Dujon, Sequence of the intron and flanking exons of the mitochondrial 21 S rRNA gene of yeast strains having different alleles at the w and RIB 1 loci. Cell (1980) 20, 185-187.

2. F. Michel, A. Jacquier and B. Dujon, Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. Biochimie, 1982, 64, 867-881.

3. F. Michel and B. Dujon, Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast-, and nuclear-encoded members. The EMBO Journal, 1983, 2, 33-38.

4. A. Jacquier and B. Dujon, The intron of the mitochondrial 21S rRNA gene: distribution in different yeast species and sequence comparison between Kluyveromyces thermotolerans and Saccharomyces cerevisiae. Mol. Gen. Gent. (1983) 192, 487-499.

5. B. Dujon and A. Jacquier, Organization of the mitochondrial 21S rRNA gene in Saccharomyces cerevisiae: mutants of the peptidyl transferase centre and nature of the omega locus in "Mitochondria 1983", Editors R. J. Schweyen, K. Wolf, F. Kaudewitz, Walter de Gruyter et Co., Berlin, New York (1983), 389-403.

6. A. Jacquier and B. Dujon, An intron encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. Cell (1985) 41, 383-394.

7. B. Dujon, G. Cottarel, L. Colleaux, M. Betermier, A. Jacquier, L. D'Auriol, F. Galibert, Mechanism of integration of an intron within a mitochondrial gene: a double strand break and the transposase function of an intron encoded protein as revealed by *in vivo* and *in vitro* assays. In Achievements and perspectives of Mitochondrial Research". Vol. II, Biogenesis, E. Quagliariello et al. Eds. Elsevier, Amsterdam (1985) pages 215-225.

8. L. Colleaux, L. D'Auriol, M. Betermier, G. Cottarel, A. Jacquier, F. Galibert, and B. Dujon, A universal code equivalent of a yeast mitochondrial intron reading frame is expressed into Escherichia coli as a specific double strand endonuclease. Cell (1986) 44, 521-533.

9. B. Dujon, L. Colleaux, A. Jacquier, F. Michel and C. Monteilhet, Mitochondrial introns as mobile genetic elements: the role of intron-encoded proteins. In "Extrachromosomal elements in lower eucaryotes", Reed B et al. Eds. (1986) Plenum Pub. Corp. 5-27.

10. F. Michel and B. Dujon, Genetic Exchanges between Bacteriophage T4 and Filamentous Fungi? Cell (1986) 46, 323.

11. L. Colleaux, L. D'Auriol, F. Galibert and B. Dujon, Recognition and cleavage site of the intron encoded omega transposase. PNAS (1988), 85, 6022-6026.

12. B. Dujon, Group I introns as mobile genetic elements, facts and mechanistic speculations: A Review. Gene (1989), 82, 91-114.

13. B. Dujon, M. Belfort, R.A. Butow, C.Jacq, C. Lemieux, P.S. Perlman, V.M. Vogt, Mobile introns: definition of terms and recommended nomenclature. Gene (1989), 82, 115-118.

14. C. Monteilhet, A. Perrin, A. Thierry, L. Colleaux, B. Dujon, Purification and Characterization of the *in vitro* activity of I-*Sce*I, a novel and highly specific endonuclease encoded by a group I intron. Nucleic Acid Research (1990), 18, 1407-1413.

15. L. Colleaux, M-R. Michel-Wolwertz, R.F. Matagne, B. Dujon - The apocytochrome b gene of Chlamydomonas smithii contains a mobile intron related to both Saccharomyces and Neurospora introns. Mol. Gen. Genet. (1990) 223, 288-296.

16. B. Dujon Des introns autonomes et mobiles. Annales de I'Institut Pasteur/ Actualites (1990) 1. 181-194.

17. A. Thierry, A. Perrin, J. Boyer, C. Fairhead, B. Dujon, B. Frey, G. Schmitz. Cleavage of yeast and bacteriophage 17 genomes at a single site using the rare cutter endonuclease I-*Sce*. I Nuc. Ac. Res. (1991) 19, 189-190.

18. A. Plessis, A. Perrin, J.E. Haber, B. Dujon, Site specific recombination determined by I-*Sce*I, a mitochondrial intron-encoded endonuclease expressed in the yeast nucleus. GENETICS (1992) 130, 451-460.

Abstracts

A1. A. Jacquier, B. Dujon. Intron recombinational insertion at the DNA level: Nature of a specific receptor site and direct role of an intron encoded protein. Cold Spring Harbor Symposium 1984.

A2. I. Colleaux, L. D'Auriol, M. Demariaux, B. Dujon, F. Galibert, and A. Jacquier, Construction of a universal code equivalent from a mitochondrial intron encoded transposase gene using oligonucleotide directed multiple mutagenesis. Colloque International de DNRS "oligonucleotids et Genetique Moleculaire" Aussois (Savoie) 8-12 January 1985.

A3. L. Colleaux, D'Auriol, M. Demariaux, B. Dujon, F. Galibert, and A. Jacquier, Expression in E. coli of a universal code equivalent of a yeast mitochondrial intron reading frame involved in the integration of an intron within a gene. Cold Spring Harbor Meeting on "Molecular Biology of Yeast", August 13-19, 1985.

A4. B. Dujon, G. Cottarel, L. Colleaux, M. Demariaux, A. Jacquier, L. D'Auriol, and F. Galibert, Mechanism of integration of an intron within a mitochondrial gene: a double strand break and the "transposase" function of an intron encoded protein as revealed by *in vivo* and *in vitro* assays. International symposium on "Achievements and Perspectives in Mitochondrial Research", Selva de Fasono (Brindisi, Italy) 26 September 1985.

A5. L. Colleaux, G. Cottarel, M. Betermier, A. Jacquier, B. Dujon, L. D'auriol, and F. Galibert, Mise en evidence de l'activite endonuclease double brin d'unc protein codee par un intron mitochondrial de levure. Forum sur la Biologie Moleculaire de la levure, Bonbannes, France 2-4 October 1985.

A6. B. Dujon, L. Colleaux, F. Michel and A. Jacquier, Mitochondrial introns as mobile genetic elements. In "Extrachromosomal elements in lower eucaryotes", Urbana, Illinois, 1-5 June 1986.

A7. L. Colleaux and B. Dujon, Activity of a mitochondrial intron encoded transposase. Yeast Genetics and Molecular Biology Meeting, Urbana, Illinois 3-6 June 1986.

A8. L. Colleaux and B. Dujon, The role of a mitochondrial intron encoded protein. XIIIth International Conference on Yeast Genetics and Molecular Biology, Banff, Alberta (Canada) 31 August - 5 September 1986.

A9. L. Colleaux, L. D'Aurio, F. Galibert and and B. Dujon, Recognition and cleavage specificity of an intron encoded transposase. 1987 Meeting on Yeast Genetics and Molecular Biology. San Francisco, CA 16-21 June 1987.

A10. A. Perrin, C. Monteilhet, L. Colleaux and B. Dujon, Biochemical activity of an intron encoded transposase of Saccharomyces cerevisiae. Cold Spring Harbor Meeting on "Molecular Biology of Mitochondria and chloroplasts" 25-30 August 1987 Cold Spring Harbor, N.Y.

A11. B. Dujon, A. Jacquier, L. Colleaux, C. Monteilhet, A. Perrin, "Les Introns autoepissables et leurs proteins" Colloque "Biologie Moleculaire de la levure: expression genetique chez Saccharomyces "organise par la Societe francaise de Microbiologie 18 January 1988 Institut Pasteur, Paris.

A12. L. Colleaux, L. D'Auriol, C. Monteilhet, F. Galibert and B. Dujon, Characterization of the biochemical activity of an intron encoded transposase. 14th International Conference on Yeast Genetics and Molecular Biology. Espoo, Finland, 7-13 August 1988.

A13. B. Dujon, A goup I intron as a mobile genetic element, Albany Conference sur "RNA: catalysis, splicing, evolution", Albany, New York, 22-25 September 1988.

A14. B. Dujon, L. Colleaux, C. Monteilhet, A. Perrin, L. D'Auriol, F. Galibert, Group I introns as mobile genetic elements: the role of intron encoded proteins and the nature of the target site. 14th Annual EMBO Symposium "Organelle genomes and the nucleus" Heidelberg, 26-29 September 1988.

A15. L. Colleaux, R. Matagne, B. Dujon, A new mobile mitochondrial intron provides evidence for genetic exchange between Neurospora and Chlamydomonas species. Cold Spring Harbor, May 1989.

A16. L. Colleaux, M.R. Michel-Wolwertz, R.F. Matagne, B. Dujon, The apoxytochrome b gene of Chlamydomonas smithii contains a mobile intron related to both Saccharomyces and Neurospora introns. Fourth International Conference

on Cell and Molecular Biology of Chlamydomonas. Madison, WI, April 1990.

A17. B. Dujon, L. Colleaux, E. Luzi, C. Monteilhet, A. Perrin, A. Plessis, I. Stroke, A. Thierry, Mobile Introns, EMBO Workshop on "Molecular Mechanisms of transposition and its control, Roscoff (France) June 1990.

A18. A. Perrin, C. Monteilhet, A. Thierry, E. Luzi, I. Stroke, L. Colleaux, B. Dujon. I-*Sce*I, a novel double strand site specific endonuclease, encoded by a mobile group I intron in Yeast. Workshop on "RecA and Related Proteins" Sacly, France 17-21 September 1990.

A19. A. Plessis, A. Perrin, B. Dujon, Site specific recombination induced by double strand endonucleases, HO and I-*Sce*I in yeast. Workshop on "RecA and Related Proteins" Saclay, France 17-21 September 1990.

A20. B. Dujon, The genetic propagation of introns 20th FEBS Meeting, Budapest, Hungary, August 1990.

A21. E. Luzi, B. Dujon, Analysis of the intron encoded site specific endonuclease I-*Sce*I by mutagenesis, Third European Congress on Cell Biology, Florence, Italy, September 1990.

A22. B. Dujon, Self splicing introns as contagious genetic elements. Journees Franco-Beiges de Pont a Mousson. October 1990.

A23. B. Frey, H. Dubler, G. Schmitz, A. Thierry, A. Perrin, J. Boyer, C. Fairhead, B. Dujon, Specific cleavage of the yeast genome at a single site using the rare cutter

endonuclease I-*Sce*I Human Genome, Frankfurt, Germany, November 1990.

A24. B. Dujon, A. Perrin, I. Stroke, E. Luzi, L. Colleaux, A. Plessis, A. Thierry, The genetic mobility of group I introns at the DNA level. Keystone Symposia Meeting on "Molecular Evolution of Introns and Other RNA elements", Taos, New Mexico, 2-8 February 1991.

A25. B. Dujon, J. Boyer, C. Fairhead, A. Perrin, A Thierry, Cartographie chez la levure. Reunion "Strategies d'etablissement des cartes geniques" Toulouse 30-31 Mai 1991.

A26. B. Dujon, A. Thierry, Nested chromosomal fragmentation using the meganuclease I-*Sce*I: a new method for the rapid mapping of the yeast genome. Elounda, Crete 15-17 Mail 1991.

A27. A. Thierry, L. Gaillon, F. Galibert, B. Dujon. The chromosome XI library: what has been accomplished, what is left. Brugge meeting 22-24 September 1991.

A28. B. Dujon, A. Thierry, Nested chromosomal fragmentation using the meganuclease I-*Sce*I: a new method for the rapid physical mapping of the eukaryotic genomes. Cold Spring Harbor 6-10 May 1992.

A29. A. Thierry, L. Gaillon, F. Galibert, B. Dujon. Yeast chromosome XI: construction of a cosmid contig. a high resolution map and sequencing progress. Cold Spring Harbor 6-10 May 1992.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
1 202 408 4000

52

In Preparation

P1. A. Thierry and B. Dujon, Nested Chromosomal Fragmentation Using the Meganuclease I-*Sce*I: Application to the physical mapping of a yeast chromosome and the direct sorting of cosmid libraries. Probably Submission to GENOMICS or EMBO J.

P2. A. Thierry, L. Colleaux and B. Dujon: Construction and Expression of a synthetic gene coding for the meganuclease I-*Sce*I. Possible submission: NAR, EMBO J.

P3. I. Stroke, V. Pelicic and B. Dujon: The evolutionarily conserved dodecapeptide motifs of intron-encoded I-*Sce*I are essential for endonuclease function. Submission to EMBO J.

P4. C. Fairhead and B. Dujon: Consequences of a double strand break induced *in vivo* in yeast at specific artificial sites, using the meganuclease I-*Sce*I. Possible submission to GENETICS, NATURE.

P5 A. Perrin, and B. Dujon: Asymetrical recognition by the I-*Sce*I endonuclease on exon and intron sequences reveals a new step in intron mobility. Possible submission: NATURE

The entire disclosure of each of these publications and abstracts is relied upon and incorporated by reference herein.

LAW OFFICES
FINNEGAN, HENDERSON
FARABOW, GARRETT
& DUNNER
1300 I STREET, N.W.
WASHINGTON, DC 20005
(202) 408-4000